Discovering Research Communities by Clustering Bibliographical Data

Fabrice Muhlenbach, Stéphane Lallich

▶ **To cite this version:**

Fabrice Muhlenbach, Stéphane Lallich. Discovering Research Communities by Clustering Bibliographical Data. IEEE/WIC/ACM International Conference on Web Intelligence, WI 2010, Aug 2010, Toronto, Canada. pp.500-507. hal-00516610

HAL Id: hal-00516610
https://hal.science/hal-00516610

Submitted on 10 Sep 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Discovering Research Communities by Clustering Bibliographical Data

Fabrice Muhlenbach
Université de Lyon
CNRS, UMR 5516, Laboratoire Hubert Curien
Université Jean Monnet, F-42000, Saint-Étienne, France
fabrice.muhlenbach@univ-st-etienne.fr

Stéphane Lallich
Université de Lyon
Laboratoire ERIC (EA 3083)
Université Lumière Lyon 2, F-69676, Bron, France
stephane.lallich@univ-lyon2.fr

***Abstract*—Today's world is characterized by the multiplicity of interconnections through many types of links between the people, that is why mining social networks appears to be an important topic. Extracting information from social networks becomes a challenging problem, particularly in the case of the discovery of community structures.**

Mining bibliographical data can be useful to find communities of researchers. In this paper we propose a formal definition to consider the similarity and dissimilarity between individuals of a social network and how a graph-based clustering method can extract research communities from the DBLP database.

***Keywords*-bibliographical data; graph-based clustering; community mining.**

I. Introduction

A social network is a social structure made of nodes representing individuals, which are connected by one or more specific types of relations, e.g., friendship, financial exchange, sexual relationships, work collaboration, etc. Extracting knowledge to understand the social relationships between individuals from this kind of data structure is a challenging data mining problem [1] because the hidden information is implicit within relationships among entities in the data, e.g., discovering organizational relations, studying the spread of disease or identifying some communities.

In this paper, we focus on the research community which can be considered as a particular social network. Nowadays we can easily access to scientific bibliographical databases –e.g., the DBLP (Digital Bibliography & Library Project) Computer Science Bibliography[1]– and this information can be useful for researchers, research organizations and institutions tasked with funding scientific research. Nevertheless it is very difficult to derive a benefit from this global information about the authors, their papers and the journals or conferences they publish in.

Clustering scientific literature to discover communities (i.e., groups of entities that share similar properties or connect to each other via certain relations) is very useful in various applications, e.g., for a researcher, to suggest potential collaborators or to propose conferences or journals where he can publish his work.

The main contributions of this paper are as follows:

- A formal definition for similarity and dissimilarity scoring over social networks (Section III).
- A clustering method using the neighborhood graph obtained with the dissimilarity scoring (Section IV).
- A graph-theoretic model for discovering research communities with DBLP database (Section V).
- Experimental results run on DBLP to discover conference communities (Section VI).

Discussions of related work on social network and research communities are provided in Section II.

II. Social Network and Research Communities

Social network analysis has emerged as an important technique in modern sociology but also in many other human or natural sciences. This analysis, related to the graph theory and the social structures, has still a long story [2] and has now moved from being a suggestive metaphor to a complete research domain with its own theoretical statements and methods, due to the arrival of the Internet (and the social networking websites) and other information and communication technologies (e.g., mobile and smart phones).

In this context, the data are by definition extremely connected from the one to the others, and to discover the knowledge hidden in this particular data structure, we have to use appropriate data mining paradigms and techniques. One important point is the discovery of community structures in the network. A community structure [3] is a group of entities that share similar properties or connect to each other with certain relations [4], i.e., there is a high density of connections within a community and a lower density between different communities.

To identify similarities between entities of a network, Jeh and Widom [5] have assumed that two objects are similar if they are related to similar objects and they have proposed an algorithm to compute the similarity scores between nodes (objects) based on the structural context in which they appear. Nevertheless the recursive computation of the score is very time consuming and this structural-context similarity score is considered by the authors as only one component of similarity which has to be combined with other domain-specific similarity measures.

[1] http://dblp.uni-trier.de/

Although many domains of "community mining" exist (e.g., friendship network, www, massive multi-player online gaming, electronic communications), we focus on the research community mining because it is important to provide a better knowledge to our professional domain.

We notice that most of the work realized in the community mining field combines unsupervised machine learning algorithms (and especially clustering like *k*-means [6]) with text mining techniques. In addition, this research is very concerned with the visualization of the obtained results, e.g., Klink et al. propose a userfriendly interface to search authors and publications and to analyse social networks on the basis of bibliographical data by combining both textual and visual browsing functionality [7].

DBconnect proposed by Zaïane et al.[4] is a navigational system based on the DBLP database which tries to reveal interesting knowledge about the research community and to recommend collaborations by combining a random walk approach with text mining techniques (the most frequent bigrams and tri-grams extracted from the paper titles).

By constructing a word association network from DBLP bibliography records, Huang et al. [8] propose to detected semantic communities and their evolutions.

Ichise et al. propose a method to discover research communities by building a network model of papers with co-citation and co-author relationships analysis and they realize a word assignment technique for the communities obtained [9], [10].

Popescul et al. realize a clustering with the citation patterns of a database to form soft clusters about the most frequently cited papers, then these soft clusters are merged by a secondary (hard) clustering algorithm (the Ward's hierarchical clustering [11] or *k*-means clustering [12]).

III. Modelization Phase

A. Matrix Formalization of Social Relationships

The matrix formalization is a key tool for social network analysis. Let a set of n individuals, denoted by I, the common use is to consider a matrix with the general term for representing the social relationships between the different individuals of I.

Depending on the properties of the social relationship analysed, this matrix can be modelized in different ways:

- Modelization 1: if we consider a mutual attraction-repulsion graph, $x_{ij} = 1$ if i is attracted to j, -1 if i rejects j and 0 in the indifference case, we will have a non-symmetric matrix where the diagonal term equals to 1.
- Modelization 2: if we consider the case of an exchange, like the email exchange, x_{ij} will be the number of emails that i has sent to j. The corresponding matrix X is a number matrix whose all line and column sums can be considered as the total number of email written by each individual. This matrix can be binarized if we do not take into account the quantities and $x_{ij} = 1$ will mean that i sends emails to j.
- Modelization 3: the social matrix can be seldom designed with reference to another set, denoted by C, e.g., an activity set. In this case, where the social matrix is denoted by Y, $y_{ij} = 1$ iff the individual i has taken part in the activity j, 0 else. The matrix Y is rectangular and its row and column sums indicate the importance of the individuals and of the activities respectively.
- Modelization 4: the social matrix X evaluates the co-activity of individuals i and j.

In this paper we try to discover research communities by analyzing the DBLP database. The most appropriate modelization in this case is the fourth where x_{ij} is the number of papers co-written by i and j. In such a matrix, the row and column sums do not have any meaning because some quantities are not uniquely recorded in the matrix, and the diagonal term x_{ii} represents the global number of activities of the individual i which is maximum for the row and the column. By considering the modelization 3, with the set of authors denoted by I and the set of activities (i.e., scientific papers) denoted by C, we will have $X = Y^T Y$ where Y is the co-activity matrix and Y^T is its transposition.

The data used in this work correspond to the relationships between international conferences (having 5 editions or more) by taking into account the fact that they share the same participants. In this case, J is the set of international conferences, I is the set of authors and x_{ij} denotes the number of authors having published at least one paper in the conference i and in the conference j.

B. Dissimilarity Matrix Construction

The dissimilarity matrix obtained by the social matrix has to take into account the nature of the social matrix values. If this matrix results from the modelization 2, it seems to be logical to evaluate the dissimilarity between the individuals i and j by the chi square distance, which eliminates the row effects and reduces the column effects.

In our case (modelization 4), with $X = Y^T Y$, the situation is more complex because the row and column sums do not have any meaning. Here, it is the diagonal term x_{ii} that will evaluate the global activity of the individual i. For taking into account these characteristics and the individual activity amount, we propose to consider the similarity between i and j by the quantity $\frac{x_{ij}}{x_{ii}}$ (in function of the activity of i) and $\frac{x_{ij}}{x_{jj}}$ (in function of the activity of j) which is synthesized with the s measure defined by:

$$s_{ij} = \frac{\left(\frac{x_{ij}}{x_{ii}} + \frac{x_{ij}}{x_{jj}}\right)}{2}.$$

The s measure is a similarity index normalized to 1, which values are between 0 and 1 (i.e., non negative) because the values $\frac{x_{ij}}{x_{ii}}$ and $\frac{x_{ij}}{x_{jj}}$ are bounded by 0 and 1; s is symmetrical

($s_{ij} = s_{ji}$); the maximal value for s is 1 iff $x_{ij} = x_{ii} = x_{jj}$, which means that the vectors representing i and j in the matrix Y are the same.

It is possible to associate to the similarity index s a value denoted by d and defined by:

$$d_{ij} = 1 - s_{ij} = 1 - \left(\frac{1}{2} \times \left(\frac{x_{ij}}{x_{ii}} + \frac{x_{ij}}{x_{jj}}\right)\right). \qquad (1)$$

The measure d is a dissimilarity index: never negative, symmetrical, d equals to 0 iff $i = j$, and the maximal value of d equals to 1 iff $x_{ij} = 0$.

C. Graph Representation

Modelizing the social relationships by a graph is not a recent idea [2]. We can associate to a square social matrix X (in order of n) a graph $G = (I, X)$. The nodes of the graph are the individuals of the set I. The nodes i and j are linked by an edge, valuated or not, oriented or not, depending on the dissimilarity matrix built from the X matrix. Some tools can be used to visualize this kind of graph [13], e.g., *Graphviz*[2] or *aiSee*[3].

When the modelization suggests a rectangular matrix Y (modelization 3), the associated graph is a bipartite graph $G = (I, C, Y)$. This graph can be visualized and synthesized by using a tool like *ZigZag* [14].

D. From the Social Network to the Neighborhood Graph

The neighborhood graphs, which are special tools of the computational geometry, can be used in many data mining tasks, and especially in the supervised machine learning [15], [16], [17], [18]. Such neighborhood structure can be for example the minimum spanning tree (MST), the relative neighborhood graph (RNG) [19], the Gabriel graph (GG) or the Delaunay triangulation. On Figure 1, we present two neighborhood graphs with the same data set.

For each graph, a specific condition is required, depending on a region of influence, to link two points with an edge. For the MST, the condition is to connect all vertices together with the minimal size of edges; for the RNG, the region of influence is a lune, the intersection of two hyperspheres centred on each pair of points (i.e., on Figure 1, the hatched area must be empty to connect the individuals α and β); and for the GG, the region of influence is an hypersphere with each pair as a diameter.

For constructing the MST or the RNG on a data set, it is not necessary to obtain the coordinates of all individuals of this data set, the only information needed is a distance matrix between all the data. By using the dissimilarity matrix with the measure d of the formula 1 proposed in Subsection III-B as a distance, we can obtain a distance matrix which can be used to construct a neighborhood graph from the initial social network.

[2]http://graphviz.org/

[3]http://www.absint.com/aisee/index

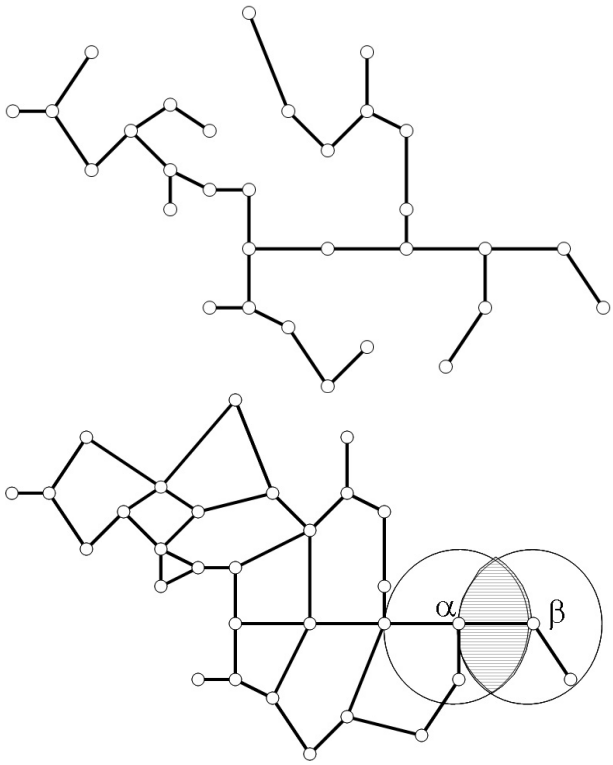

Figure 1. Two neighborhood graphs: MST (on the top) and RNG (below).

IV. Graph-Based Clustering (*GBC*)

A. Presentation of the Clustering Method

The algorithm *GBC* that we have proposed in a general data mining clustering task [20] can be easily adapted for the discovery of communities in a social network with three main advantages. First, *GBC* obtains good results when the data are well-structured: it detects easily the well-formed clusters and the outliers, whether the cluster shapes are convex or not, whether the cluster sizes are homogeneous or not. Second, the main advantage of *GBC* is that the method does not need any parameter to perform on a data set. It is a considerable improvement compared with other clustering techniques developed in the data mining literature. Third, when they are some outliers in a data set, *GBC* can automatically find them (they are detected as singletons).

B. GBC, the Clustering Algorithm

GBC is conducted in 2 phases. The first phase (in 10 steps) consists in doing a list of μ values which will be used in the formula 2 to detect the appropriate number of clusters and is conducted as written on Table I.

Notice that we can equally use any kind of connected neighborhood graph on the step 1 of the algorithm. In the experiments performed on [20], we did not find significant differences in the results, but following [21], we recommend

1	construction of a neighborhood graph NG
2	descent sorting (by size) of the edge set E of NG
3	initialization step: $k \leftarrow 2$, $\Sigma_k \leftarrow 0$, and $n_e \leftarrow 1$
4	cutting the edge $e_{\max} \in E$ of the (sub-)graph with the higher value
5	adding the size of $e_{\max}$ to Σ_k, the sum of the cut edges at the level k
6	testing if the (sub-)graph with the edges $E - \{e_{\max}\}$ is still connected
7	if the graph is still connected, increasing the number of edges: $n_e \leftarrow n_e + 1$
8	if the graph is not connected, modifying: $k \leftarrow k + 1$ (new level for having k clusters) $n_e \leftarrow 1$ (re-initialization of the number of cut edges) $\mu_k \leftarrow \Sigma_k / n_e$ (μ_k is the average of the sizes of the cut edges) $\Sigma_k \leftarrow 0$ (re-initialization of the sum of cut edge sizes)
9	$E \leftarrow E - \{e_{\max}\}$ (remove the bigger edge from the set E)
10	back to step 4 by using the next maximal edge $e_{\max} \in E$ while $k < n$

Table I
GBC ALGORITHM

the RNG of Toussaint [19] which is a structure that overcomes some problems encountered with the MST.

After this first phase, we can calculate the δ values for each level k as follows:

$$\delta_k = \frac{\mu_k - \mu_{k+1}}{\mu_k + \mu_{k+1}}, \forall k = 2, \ldots, n-1 \ . \qquad (2)$$

The maximal value of δ_k is used to select $k^\star$, the *ideal* number of clusters in the data set. The second phase is similar to the first one, but the loop runs until $k = k^\star$ in the step 10 (instead of $k < n$).

C. Illustration: Discovering Laboratory Communities

To illustrate how the proposed model can detect research communities and how these communities can evolve, consider a simple case of 4 researchers referred to as A, B, C and D, see Table II.

Imagine that A is a PhD student at step t_1 and works mostly with his director, who is B. We assume that A will mainly publish with his director B and that all of his publications will be written in collaboration with him, so that $x_{AA} = x_{AB}$. Nevertheless the director B has published many other papers without his PhD student, so $x_{BB} > x_{AB}$. It is thought that B has published some papers with other researchers of his laboratory, for example C, who has also published one paper with A.

At step t_2, A has defended his PhD thesis and has been recruited for an assistant professor position at the laboratory of D. We assume that A (or any member of his previous laboratory) has never worked before with D, but A will publish now with people from his new laboratory, and mainly with D.

At step t_3, A has become an associate professor of the same laboratory than D. They are now many publications written together by A and D.

t_1	A	B	C	D
A	3	3	1	0
B	3	20	10	0
C	1	10	15	0
D	0	0	0	25
t_2	A	B	C	D
A	8	3	2	3
B	3	26	12	0
C	2	12	18	0
D	3	0	0	32
t_3	A	B	C	D
A	16	3	2	10
B	3	28	14	0
C	2	14	21	0
D	10	0	0	39

Table II
CO-WRITING VALUES AT TIMES t_1, t_2 AND t_3.

The dissimilarity values obtained with the co-writing values of the Table II are presented on Table III. On this table, we can see at t_1 that the maximal distance is obtained for a connection between D and someone else due to the lack of articles written by people from different laboratories. At t_2 the dissimilarity values between A and B or D are growing, and the value between A and D is decreasing, due to the new collaborations between A and D. This phenomenon is emphasized at t_3.

t_1	A	B	C	D
A	0.0000	0.4250	0.8000	1.0000
B	0.4250	0.0000	0.4167	1.0000
C	0.8000	0.4167	0.0000	1.0000
D	1.0000	1.0000	1.0000	0.0000
t_2	A	B	C	D
A	0.0000	0.7548	0.8194	0.7656
B	0.7548	0.0000	0.4359	1.0000
C	0.8194	0.4359	0.0000	1.0000
D	0.7656	1.0000	1.0000	0.0000
t_3	A	B	C	D
A	0.0000	0.8527	0.8899	0.5593
B	0.8527	0.0000	0.4167	1.0000
C	0.8899	0.4167	0.0000	1.0000
D	0.5593	1.0000	1.0000	0.0000

Table III
CO-WRITING DISSIMILARITY VALUES AT t_1, t_2 AND t_3.

On Table IV we present the general results obtained with the clustering method *GBC* on the dissimilarity values of the previous table (following the algorithm presented on Table I, n_e is the number of edges cut to obtain k sub-graphs from $k-1$ sub-graphs, Σ is the sum of the cut edge sizes, μ is the average cut edge size and δ is the value computed with the formula 2). At t_1, the method indicates that δ will be maximal for $k = 2$, which means that it is relevant to have

two clusters of 3 elements and 1 element (for $\{A, B, C\}$ and $\{D\}$ respectively). At t_2, the method proposes three clusters: one with two elements (with $\{B, C\}$), the two others are singletons ($\{A\}$ and $\{D\}$). At t_3, δ will be maximal for $k = 2$, with two clusters of two elements ($\{A, D\}$ in one cluster, $\{B, C\}$ in the other cluster).

t_1	n_e	$\Sigma(\%)$	$\mu(\%)$	$\delta(\%)$	cluster sizes
$k=1$	0	0.00	0.00	—	4
$k=2$	3	78.09	26.03	**40.35**	3 1
$k=3$	1	11.06	11.06	0.99	1 2 1
$k=4$	1	10.85	10.85	—	1 1 1 1
t_2	n_e	$\Sigma(\%)$	$\mu(\%)$	$\delta(\%)$	cluster sizes
$k=1$	0	0.00	0.00	—	4
$k=2$	1	39.14	39.14	0.71	3 1
$k=3$	1	38.58	38.58	**26.78**	1 2 1
$k=4$	1	22.28	22.28	—	1 1 1 1
t_3	n_e	$\Sigma(\%)$	$\mu(\%)$	$\delta(\%)$	cluster sizes
$k=1$	0	0.00	0.00	—	4
$k=2$	1	46.63	46.63	**20.78**	2 2
$k=3$	1	30.59	30.59	14.61	1 1 2
$k=4$	1	22.79	22.79	—	1 1 1 1

Table IV
GENERAL RESULTS OBTAINED WITH *GBC* AT t_1, t_2 AND t_3.

The Figure 2 synthesizes this illustrative example for the steps t_1, t_2 and t_3 with three collaboration graphs (like the *Erdös collaboration graph* where two mathematicians are joined by an edge whenever they co-authored a paper together with possibly other co-authors present). The authors are the nodes of the graph. The publication numbers of each author are indicated in the brackets, and the numbers of papers co-written by the authors are indicated on the edges. At t_1, the social network connects the individuals A, B and C only. The clustering method *GBC* gathers A, B and C together, and this cluster corresponds to a given laboratory. At t_2, the individual A is out of the cluster $\{B, C\}$, but does not belong yet to the same cluster of the individual D. At t_3, the individuals A and D are in the same cluster.

Finally, this example illustrates how the method can automatically discover clusters corresponding to communities of authors publishing together, like people belonging to the same laboratories.

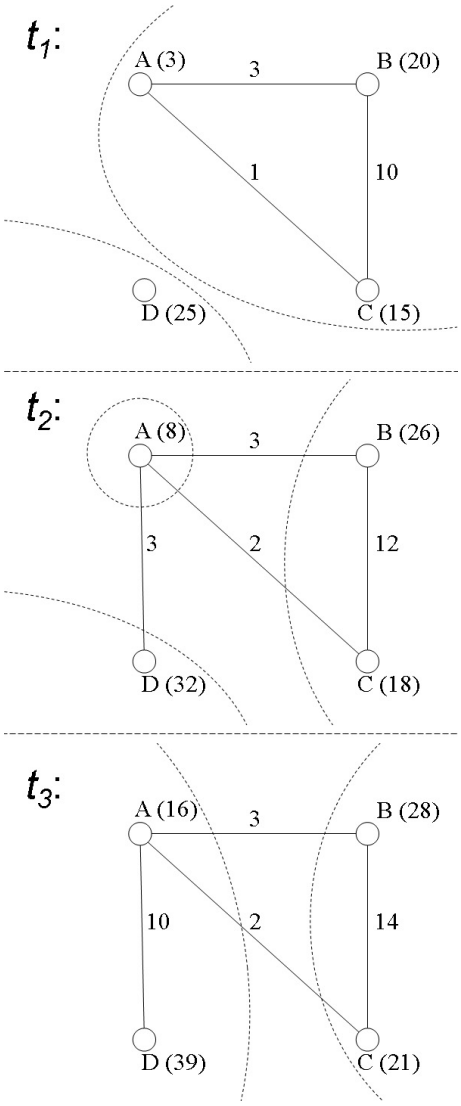

Figure 2. Research community clusters obtained at t_1, t_2 and t_3.

V. DISCOVERING RESEARCH COMMUNITIES WITH DBLP DATABASE

Like CiteSeer or Google-Scholar, DBLP is a huge digital library which provides access to computer science publications. The data quality of this collection is very important although some problems related to the person names can be found [22]. DBLP now lists more than 1.3 million publications which are archived on a XML file that can be downloaded on the Internet[4]. By parsing this XML file, we can research information by authors, conferences, journals, series or books collections. Unfortunately they are not topic elements in this file.

Text mining techniques apart, the most common task to do with the DBLP bibliographical data set is to extract the relation between authors and conferences or journals.

The conference-author matrix can be used to identify easily the most prolific authors or the most important (in term of publication quantities) conferences.

By taking into account the papers co-written by authors, we can extract the (Erdös) collaboration graph from which we can extract some research communities of authors publishing together, like people belonging to the same laboratories, as illustrated on Subsection IV-C.

We propose to extract information from DBLP database to detect clusters of conferences. On Figure 3 we present

[4] http://dblp.uni-trier.de/xml/

the model used to do this task:

- From the `dblp.xml` file we extract the list of all authors having published in (at least) one conference: we obtain 578,330 author names.
- From the `dblp.xml` file we extract the list of all conferences. From this list, we keep only the international conference (with "International" explicitly in the conference title). We keep only the "historical" conferences (at least 5 different editions) and we obtain 574 conferences.
- For each author, we list all the conferences where he has published an article.
- From the author-conference list, we can obtain the square matrix X with the values x_{ij} corresponding to the number of different authors having published at the same time in the conference i and in the conference j.
- With the square matrix X, we can calculate the dissimilarity measure d (formula 1), build a neighborhood graph and use the clustering method *GBC*.

The experimental results obtained with this method are presented in the next section.

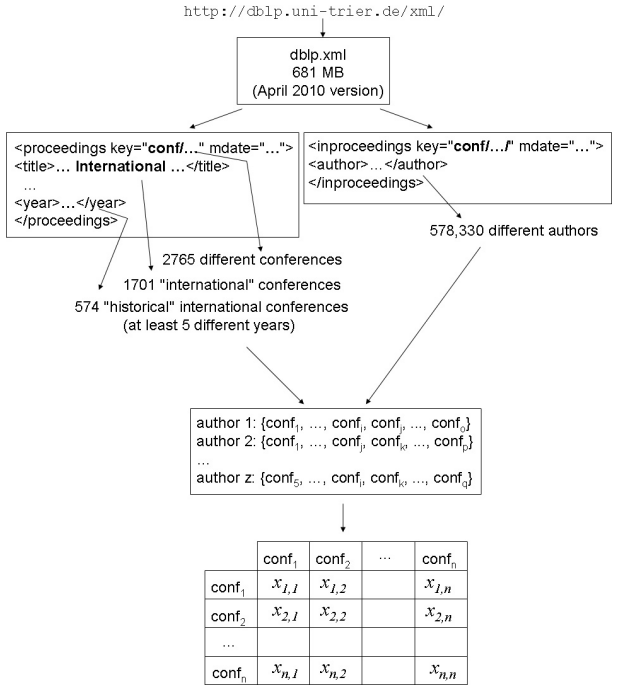

Figure 3. Model to extract the similarity values between conferences from the DBLP database.

VI. Experimental Results on the Bibliographical Database

A. Bottom-Up Clustering

On Table V, we present the values obtained on the 574 conferences for $k^\star$, the best number of clusters proposed by the clustering method *GBC*.

$k^\star$	$\delta(\%)$
572	3.55021
573	3.43267
550	1.10274
569	1.05070
566	0.66557
563	0.50048
555	0.35267
540	0.33291
565	0.32914
561	0.32136

Table V
First tenth experimental values of $k^\star$ for the maximal values of δ.

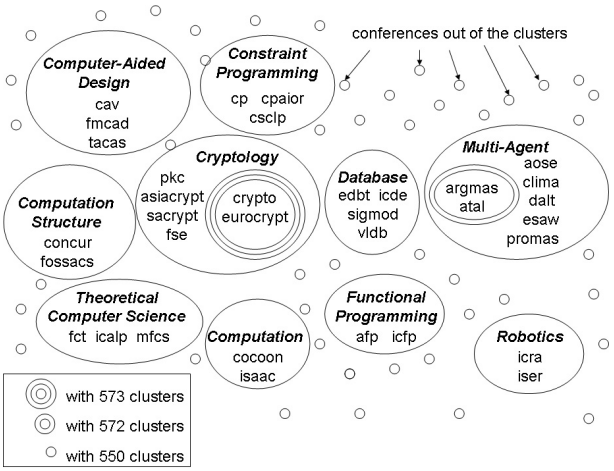

Figure 4. Conference clusters and topics obtained for 550, 572 and 573 clusters.

These first results seem to be a little disappointing: the $k^\star$ values proposed by *GBC* are close to $n = 574$, i.e., all the data can be considered as singletons. Nevertheless if we consider the three first values of $k^\star$, we can obtain 572, 573 or 550 clusters, which are the first steps of an agglomerative process of clustering, linking the most similar conferences. The results are presented on Figure 4.

With 573 clusters, in addition of 572 singletons, they are two conferences linked on the same cluster, which are two conferences about cryptology.

With 572 clusters, there is another cluster with two conferences about multi-agent.

With 550 clusters, we will find some interesting clusters of conferences (in this example, the topics are given manually by looking for the most similar words in the conference names[5]).

[5]We can find information about the conferences on DBLP website simply by adding the conference code (e.g., "crypto") at the end of the URL: `http://www.informatik.uni-trier.de/~ley/db/conf/`

B. Top-Down Clustering

The divisive (or top-down) process is the most natural way of clustering of the method *GBC* because it processes by cutting progressively a connected neighborhood graph. On Table VI, we present the proposed values for $k^\star$, the best number of clusters, after having removed the values too close to n which are handled in the previous subsection.

$k^\star$	$\delta(\%)$
2	0.27267
10	0.23451
9	0.20177
4	0.13077
417	0.11444
3	0.09508
22	0.08253
250	0.07863
117	0.07857
69	0.07194
159	0.06902
23	0.06695

Table VI
FIRST TENTH EXPERIMENTAL VALUES OF $k^\star$ FOR THE MAXIMAL VALUES OF δ OBTAINED ON THE DATA SET WHEN $k^\star \ll n$.

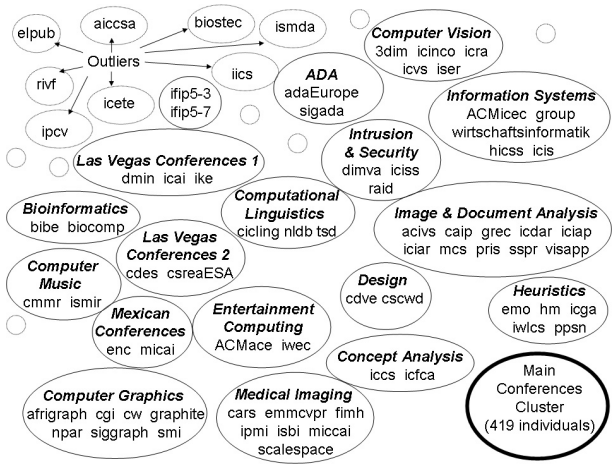

Figure 5. Conference clusters obtained for 2 to 117 clusters.

On Figure 5, we present the clusters obtained with a divisive procedure, separating clusters and outliers from the main conference cluster (from the top left to the bottom right of the figure).

The partition obtained with 2 clusters is a singleton and a cluster with the 573 other conferences. Actually, the cluster with only one conference (*ELPUB*, the International Conference on Electronic Publishing) can be considered as an outlier. By definition, an outlier is an observation that deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism [23]. We notice that *GBC* is very sensitive to the outliers because an outlier will be far from the other data in the representation space, and it will be detected as an independent cluster.

The *ELPUB* conference has some interesting properties: in the co-occurrence matrix X it is the unique conference where $c_{ii} = 745$ is as small as $\sum_{j=1}^{n} c_{ij} = 750$ with $i \neq j$: actually, they are many *ELPUB*-authors who have published their work only in this conference, and most of them have published only one article in their scientific life.

Moreover, it was sometimes difficult to find manually the topic of a conference cluster, e.g., the conferences DMIN (International Conference on Data Mining), IC-AI (International Conference on Artificial Intelligence) and IKE (International Conference on Information and Knowledge Engineering) belong to the same cluster but the related topics of these conferences seem to differ notably (e.g., DMIN does not belong to the other data mining conference cluster composed by ICDM, KDD, KDID and SDM).

We have then discovered that all these conferences put together in a cluster by *GBC* without having an easily identifiable topic are always organized at the same place (e.g., Las Vegas, USA, or Mexico). Even if the conferences selected for the experiment are considered as "international conferences", we can imagine that some communities of researchers will preferably go to some specific places (local or attractive places) or if they will attend to a conference at a given place they will try to publish scientific papers in other conferences those will be held at the same place.

VII. CONCLUSION AND FUTURE WORK

This paper deals with the community mining. It considers different kind of social relationships and proposes for each of them a well-suited modelization. In the case of a conference-author matrix, an adapted dissimilarity index is build. From this dissimilarity index, a neighborhood graph is constructed. *GBC* clustering method allows to discover research communities in the DBLP database.

Heartened by the first results obtained on the conference clustering, we take under consideration the following perspectives:

- Studying the evolution of conference communities by taking into account the publication dates, seeing if the conference participants are the same or change mostly from an year to another.
- Adding the journals to the conferences and finding some behaviors and trends, e.g., after having published a paper in a given conference, in which journal the authors will publish an extended version of their work.
- Proposing to find automatically the most relevant cluster for a given individual (a conference or an author) in the way of having a data-driven method to detect the best value $k^\star$ for the clustering method *GBC*.
- And finally finding automatically the properties of a given cluster (same topic by analysing the most frequent words of the conference names, or same place).

REFERENCES

[1] Q. Yang and X. Wu, "10 challenging problems in data mining research," *International Journal of Information Technology and Decision Making*, vol. 5, no. 4, pp. 597–604, 2006.

[2] C. Flament, *Théorie des graphes et structures sociales*. Paris: Mouton, Gauthier-Villars, 1965.

[3] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[4] O. R. Zaïane, J. Chen, and R. Goebel, "Mining research communities in bibliographical data," in *Advances in Web Mining and Web Usage Analysis, 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007, and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007, San Jose, CA, USA, August 12-15, 2007. Revised Papers*, ser. Lecture Notes in Computer Science, H. Zhang, M. Spiliopoulou, B. Mobasher, C. L. Giles, A. McCallum, O. Nasraoui, J. Srivastava, and J. Yen, Eds., vol. 5439. Springer, 2007, pp. 59–76.

[5] G. Jeh and J. Widom, "SimRank: a measure of structural-context similarity," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*. ACM, 2002, pp. 538–543.

[6] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1967, vol. 1, pp. 281–297.

[7] S. Klink, P. Reuther, A. Weber, B. Walter, and M. Ley, "Analysing social networks within bibliographical data," in *Database and Expert Systems Applications, 17th International Conference, DEXA 2006, Kraków, Poland, September 4-8, 2006, Proceedings*, ser. Lecture Notes in Computer Science, S. Bressan, J. Küng, and R. Wagner, Eds., vol. 4080. Springer, 2006, pp. 234–243.

[8] Z. Huang, Y. Yan, Y. Qiu, and S. Qiao, "Exploring emergent semantic communities from DBLP bibliography database," in *2009 International Conference on Advances in Social Network Analysis and Mining, ASONAM 2009, 20-22 July 2009, Athens, Greece*, N. Memon and R. Alhajj, Eds. IEEE Computer Society, 2009, pp. 219–224.

[9] R. Ichise, H. Takeda, and K. Ueyama, "Community mining tool using bibliography data," in *9th International Conference on Information Visualisation, IV 2005, 6-8 July 2005, London, UK*. IEEE Computer Society, 2005, pp. 953–958.

[10] R. Ichise, H. Takeda, and T. Muraki, "Research community mining with topic identification," in *10th International Conference on Information Visualisation, IV 2006, 5-7 July 2006, London, UK*. IEEE Computer Society, 2006, pp. 276–281.

[11] A. Popescul, G. W. Flake, S. Lawrence, L. H. Ungar, and C. L. Giles, "Clustering and identifying temporal trends in document databases," in *Proceedings of IEEE Advances in Digital Libraries 2000 (ADL 2000), 22-24 May 2000, Washington, DC*. IEEE Computer Society, 2000, pp. 173–182.

[12] A. Popescul and L. H. Ungar, "Cluster-based concept invention for statistical relational learning," in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004*, W. Kim, R. Kohavi, J. Gehrke, and W. DuMouchel, Eds. ACM, 2004, pp. 665–670.

[13] I. Herman, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[14] S. Lallich, "ZigZag, a new clustering algorithm to analyze categorical variable cross-classification tables," in *Principles of Data Mining and Knowledge Discovery, Third European Conference, PKDD '99, Prague, Czech Republic, September 15-18, 1999, Proceedings*, ser. Lecture Notes in Computer Science, J. M. Zytkow and J. Rauch, Eds., vol. 1704. Springer, 1999, pp. 398–405.

[15] F. Muhlenbach and R. Rakotomalala, "Multivariate supervised discretization, a neighborhood graph approach," in *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, V. Kumar, S. Tsumoto, N. Zhong, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2002, pp. 314–321.

[16] F. Muhlenbach, S. Lallich, and D. A. Zighed, "Identifying and handling mislabelled instances," *Journal of Intelligent Information Systems (JIIS)*, vol. 22, no. 1, pp. 89–109, 2004.

[17] D. A. Zighed, S. Lallich, and F. Muhlenbach, "A statistical approach to class separability," *Applied Stochastic Models in Business and Industry*, vol. 22, no. 2, pp. 187–197, 2005.

[18] G. T. Toussaint, "Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining," *International Journal of Computational Geometry and Applications (IJCGA)*, vol. 15, no. 2, pp. 101–150, 2005.

[19] ——, "The relative neighbourhood graph of a finite planar set," *Pattern Recognition*, vol. 12, no. 4, pp. 261–268, 1980.

[20] F. Muhlenbach and S. Lallich, "A new clustering algorithm based on regions of influence with self-detection of the best number of clusters," in *ICDM 2009, The Ninth IEEE International Conference on Data Mining, Miami, Florida, USA, 6-9 December 2009*, W. Wang, H. Kargupta, S. Ranka, P. S. Yu, and X. Wu, Eds. IEEE Computer Society, 2009, pp. 884–889.

[21] R. Urquhart, "Graph theoretical clustering based on limited neighbourhood sets," *Pattern Recognition*, vol. 15, no. 3, pp. 173–187, 1982.

[22] M. Ley and P. Reuther, "Maintaining an online bibliographical database: The problem of data quality," in *Extraction et gestion des connaissances (EGC'2006), Actes des sixièmes journées Extraction et Gestion des Connaissances, Lille, France, 17-20 janvier 2006, 2 Volumes*, ser. Revue des Nouvelles Technologies de l'Information, G. Ritschard and C. Djeraba, Eds., vol. RNTI-E-6. Cépaduès-Éditions, 2006, pp. 5–10.

[23] D. M. Hawkins, *The Identification of Outliers*. London: Chapman and Hall, 1980.